Noname manuscript No.
(will be inserted by the editor)

Regularised PCA to denoise and visualise data

Marie Verbanck · Julie Josse · François Husson

Received: date / Accepted: date

Abstract Principal component analysis (PCA) is a well-established method commonly used to explore and visualise data. A classical PCA model is the fixed effect model where data are generated as a fixed structure of low rank corrupted by noise. Under this model, PCA does not provide the best recovery of the underlying signal in terms of mean squared error. Following the same principle as in ridge regression, we propose a regularised version of PCA that boils down to threshold the singular values. Each singular value is multiplied by a term which can be seen as the ratio of the signal variance over the total variance of the associated dimension. The regularised term is analytically derived using asymptotic results and can also be justified from a Bayesian treatment of the model. Regularised PCA provides promising results in terms of the recovery of the true signal and the graphical outputs in comparison with classical PCA and with a soft thresholding estimation strategy. The gap between PCA and regularised PCA is all the more important that data are noisy.

Keywords principal component analysis · singular value thresholding · regularised PCA · fixed effect model · visualisation · denoising.

M. Verbanck
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-54-91
E-mail: marie.verbanck@agrocampus-ouest.fr

J. Josse
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-58-74
E-mail: julie.josse@agrocampus-ouest.fr

F. Husson
Applied mathematics department, Agrocampus Ouest
Tel.: +332-23-48-58-86
E-mail: francois.husson@agrocampus-ouest.fr

1 Introduction

Principal component analysis (PCA) is a well-established method which can be used to explore and visualise data. PCA can be presented from a geometrical or a model point of view. A classical PCA model is the fixed effect one (Caussinus, 1986) where a data matrix of n individuals and p variables (considered to be centered) is generated as a fixed structure having a low rank representation corrupted by noise:

$$\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \varepsilon_{n\times p} \tag{1}$$

$$x_{ij} = \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}\ ,\ \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where d_s is the s^{th} eigenvalue of the matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ (n times the true covariance matrix), $\mathbf{r}_s =$

$\{r_{1s}, ..., r_{js}, ..., r_{ps}\}$ is the associated eigenvector and $\mathbf{q}_s = \{q_{1s}, ..., q_{is}, ..., q_{ns}\}$ is the s^{th} eigenvector of the matrix $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ (n times the true inner-product matrix). Under this model, the individuals have different expectations and randomness is only due to the error term. This is in agreement with situations where PCA is performed on data where the individuals are objects of interest and are not a random sample drawn from a population of individuals. Such situations frequently arise in practice. For instance, in sensory analysis, individuals can be products, such as chocolates, and variables can be sensory descriptors, such as bitterness, sweetness, etc. The aim is to study these specific products and not others (they are not interchangeable). It thus makes sense to estimate the individual parameters ($\mathbf{q}_s$) and to study the graphical representation of the individuals as well as the representation of the variables. In addition, let us point out that the inferential framework associated with this model is not usual. Indeed the number of parameters increases when the number of individuals increases. Consequently, in this model, the asymptotic results are obtained considering the noise variance tends to 0.

The maximum likelihood estimators of this model correspond to the least squares ones, *i.e.* the usual PCA solution. More precisely, PCA finds a matrix $\hat{\mathbf{X}}_{n\times p}$, of low rank S, which minimises $\|\mathbf{X}-\hat{\mathbf{X}}\|^2$ with $\|\bullet\|$ the Frobenius norm. The solution is given by the singular value decomposition (SVD) of $\mathbf{X}$:

$$\hat{x}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} \tag{2}$$

where λ_s is the s^{th} eigenvalue of $\mathbf{X}'\mathbf{X}$, $\mathbf{u}_s = \{u_{1s}, ..., u_{is}, ..., u_{ns}\}$ the s^{th} left singular vector and $\mathbf{v}_s = \{v_{1s}, ..., v_{js}, ..., v_{ps}\}$ the s^{th} right singular vector.

It is established, in regression for instance, that the maximum likelihood estimators are not necessarily the best ones in terms of mean squared error (MSE). However, shrinkage estimators, although biased, have smaller variance which may reduce the MSE. We follow this approach and propose a regularised version of PCA in order to get a better estimate of the underlying structure $\tilde{\mathbf{X}}$ and graphical representations which are as close as possible to the representations that would be obtained from the signal only. As we will show later, the approach boils down to threshold the singular values with a different amount of shrinkage for each singular value. The shrinkage term will be analytically derived.

Singular value thresholding is a popular estimation strategy to recover a low rank signal from noisy data. The most classical approach consists in using soft thresholding. It boils down to threshold each singular value with a same amount of shrinkage usually determined by cross-validation. However, recently, Candès et al (2012) proposed to determine the threshold level without resorting to a computational method by minimising an estimate of the MSE namely a Stein's Unbiased Risk Estimate (SURE). We will confront our approach to the SURE method.

In addition, it is worth quoting the work of Takane and Hwang (2006) and Hwang et al (2009) who proposed, in an exploratory framework, a regularised version of multiple correspondence analysis (MCA). Their aim is to get parameters and graphical representations which are as close as possible to the "true" ones in terms of MSE. Their approach is thus close to our proposal but there are three main differences. Firstly, MCA is classically applied to survey data, where a sample of individuals is studied. Secondly, they do not consider any underlying model. Thirdly, the shrinkage term is determined via cross-validation.

In this paper, we derive the shrinkage terms by minimising the mean squared error and define regularised PCA (rPCA) in Section 2. We also show that rPCA can be presented from a Bayesian treatment of the fixed effect model (1). Section 3 shows the efficiency of regularisation through a simulation study in which rPCA is set against classical PCA and the SURE method. The performance of rPCA is illustrated through the recovery of the signal

and the graphical outputs (individual and variable representations). Finally, rPCA is performed on a real microarray data set in Section 4.

2 Regularised PCA

2.1 MSE point of view

2.1.1 Minimising the MSE

PCA provides an estimator $\hat{\mathbf{X}}$ which is as close as possible to $\mathbf{X}$ in the least squares sense. However, assuming model (1), the objective is to get an estimator as close as possible to the unknown signal $\tilde{\mathbf{X}}$. To achieve such a goal, the same principle as in ridge regression is followed. We look for a shrinkage version of the maximum likelihood estimator which is as close as possible to the true structure. More precisely, we look for shrinkage terms $(\phi_s)_{s=1,\ldots,\min(n-1,p)}$ that minimise:

$$\mathrm{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{\min(n-1,p)} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}\right)^2\right)$$
$$\text{with } \hat{x}_{ij}^{(s)} = \sqrt{\lambda_s} u_{is} v_{js}$$
$$= \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2 + \sum_{s=S+1}^{\min(n-1,p)} \left(\phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2\right)$$

According to equation (1), for all $s \geq S+1$, $\tilde{x}_{ij}^{(s)} = 0$, therefore $\phi_{S+1} = \ldots = \phi_{\min(n-1,p)} = 0$. Thus, the MSE can be written as:

$$\mathrm{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s \hat{x}_{ij}^{(s)} - \tilde{x}_{ij}^{(s)}\right)^2\right)$$

Using the orthogonality constraints, for all $s \neq s'$, $\sum_i u_{is} u_{is'} = \sum_j v_{js} v_{js'} = 0$, the MSE can be simplified as follows:

$$\mathrm{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s^2 \lambda_s u_{is}^2 v_{js}^2 - 2\tilde{x}_{ij} \sum_{s=1}^{S} \phi_s \sqrt{\lambda_s} u_{is} v_{js} + (\tilde{x}_{ij})^2\right)\right) \quad (3)$$

Equation (3) is differentiated with respect to ϕ_s to get:

$$\phi_s = \frac{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)}\right) \tilde{x}_{ij}}{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)2}\right)} = \frac{\sum_{i,j} \mathbb{E}\left(\hat{x}_{ij}^{(s)}\right) \tilde{x}_{ij}}{\sum_{i,j}\left(\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right) + \left(\mathbb{E}\left(\hat{x}_{ij}^{(s)}\right)\right)^2\right)}$$

Based on the results of Denis and Pázman (1999) and Denis and Gower (1996), we can state that the PCA estimator is asymptotically unbiased $\mathbb{E}(\hat{x}_{ij}^{(s)}) = \tilde{x}_{ij}^{(s)}$. More precisely, they studied nonlinear regression models with constraints and focused on bilinear models in an analysis of variance framework with two factors, called biadditive models. Such models are defined as follow:

$$y_{ij} = \mu + \alpha_i + \beta_j + \sum_{s=1}^{S} \gamma_{is} \delta_{js} + \varepsilon_{ij} \quad (4)$$
$$\text{with } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where μ is the grand mean, $(\alpha_i)_{i=1,\ldots,I}$ and $(\beta_j)_{j=1,\ldots,J}$ correspond to the main effect parameters and $(\gamma_{is}\delta_{js})_{s=1,\ldots,S}$ models the interaction. The least square estimates of the multiplicative terms are given by the singular value decomposition of the residual matrix of the model without interaction. Using the Jacobians and the Hessians of the response defined by Denis and Gower (1994) and recently in Papadopoulo and Lourakis (2000), Denis and Gower (1996) derived the asymptotic bias of the response and showed that the response estimator is approximately unbiased. In the PCA framework, it leads to state that asymptotically $\mathbb{E}(\hat{x}_{ij}) = \tilde{x}_{ij}$ and for each dimension s, $\mathbb{E}\left(\hat{x}_{ij}^{(s)}\right) = \tilde{x}_{ij}^{(s)}$. In addition, we can consider

that the variance of $\hat{x}_{ij}$ is equal to the noise variance. Therefore, we estimate $\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right)$ by the average variance, that is $\mathbb{V}\left(\hat{x}_{ij}^{(s)}\right) = \frac{1}{\min(n-1;p)}\sigma^2$.

Consequently ϕ_s can be approximated by:

$$\phi_s = \frac{\sum_{i,j} \tilde{x}_{ij}^{(s)} \tilde{x}_{ij}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + \tilde{x}_{ij}^{(s)2}\right)}$$

As for all $s \neq s'$, the dimensions s and s' of $\tilde{\mathbf{X}}$ are orthogonal, ϕ_s can be written as:

$$\phi_s = \frac{\sum_{i,j} \tilde{x}_{ij}^{(s)2}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + \tilde{x}_{ij}^{(s)2}\right)}$$

Based on equation (1), $\sum_{i,j}(\tilde{x}_{ij}^{(s)})^2$ is equal to d_s the variance of the s^{th} dimension of the signal. ϕ_s is then equal to:

$$\phi_s = \begin{cases} \dfrac{d_s}{\frac{np}{min\{p,n-1\}}\sigma^2 + d_s} \forall s = 1, ..., S \\ 0 \qquad\qquad\qquad\qquad \text{else wise} \end{cases} \tag{5}$$

The form of the shrinkage term is appealing since it corresponds to the ratio of the variance of the signal over the total variance (signal plus noise) for the s^{th} dimension.

Remark: This term is obtained using asymptotic results ($\sigma^2 \to 0$), however Denis and Pázman (1999) verified that even in situations far from the asymptotic, these approximations are reasonable. We also checked this assumption through simulations.

Remark: Models such as model (4) *are also known as Additive Main effects and Multiplicative Interaction models (AMMI). They are often used to analyse genotype-environment data in plant breading framework. Considering a random version of such models, Cornelius and Crossa (1999) came up with a regularisation term which is close to ours. It allows them to improve the prediction of the yield of genotypes in relation with environments.*

2.1.2 Definition of the regularised PCA

The shrinkage factor (5) depends on unknown quantities. We estimate them by plug-in. The total variance of the s^{th} dimension is estimated by the eigenvalue λ_s. The signal variance of the s^{th} dimension is estimated by the estimated total variance of the s^{th} dimension minus an estimate of the noise variance on the s^{th} dimension. Consequently, ϕ_s is estimated by $\hat{\phi}_s = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$. Regularised PCA (rPCA) is thus defined as:

$$\begin{aligned} \hat{x}_{ij}^{\text{rPCA}} &= \sum_{s=1}^{S}\left(\frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}\right)\sqrt{\lambda_s} u_{is} v_{js} \\ &= \sum_{s=1}^{S}\left(\sqrt{\lambda_s} - \frac{\frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\sqrt{\lambda_s}}\right) u_{is} v_{js} \end{aligned}$$

rPCA boils down to threshold the first S singular values. It can be interpreted as a compromise between hard and soft thresholding. Hard thresholding consists in selecting a certain number of dimensions S which corresponds to classical PCA (equation 2) whereas soft thresholding consists in thresholding all singular values with the same amount of shrinkage. In rPCA, the s^{th} singular value is less shrunk than the $(s+1)^{th}$ one. This can be interpreted as granting a greater weight to the first dimensions. This behaviour seems satisfactory. Indeed, the first dimensions can be considered as more stable and trustworthy than the last ones. The regularisation procedure is more relying on what is less variable. When $\hat{\sigma}^2$ is small, $\hat{\phi}_s$ is close to 1 and rPCA comes down to perform PCA. When $\hat{\sigma}^2$ is high, $\hat{\phi}_s$ is close to 0 and the values of $\hat{\mathbf{X}}^{\text{rPCA}}$ are close to 0 which corresponds to the average of the variables (because $\hat{\mathbf{X}}^{\text{rPCA}}$ is centered). From a geometrical point of view rPCA leads to bring the individuals closer to the centre of gravity.

The regularisation procedure requires to estimate the residual variance σ^2. As the maximum likelihood estimator is biased, another estimator corresponds to the ratio of the residual sum of squares divided by the number of data minus the number of independent parameters.

These latter are equal to $p+\left((nS-S)-\frac{S(S+1)}{2}\right)+\left(pS-\frac{S(S+1)}{2}-S\right)$, *i.e.* p parameters for the centering, $\left((nS-S)-\frac{S(S+1)}{2}\right)$ for the centered and orthonormal left singular vectors and $\left(pS-\frac{S(S+1)}{2}-S\right)$ for the orthonormal right singular vectors. This number of parameters can also be calculated as the trace of the projection matrix involved in PCA (Candès and Tao, 2009; Josse and Husson, 2011). Therefore, the residual variance is estimated as:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\|\mathbf{X}-\hat{\mathbf{X}}\|^2}{np-p-nS-pS+S^2+S} \\ &= \frac{\sum_{s=S+1}^{\min(n-1;p)}\lambda_s}{np-p-nS-pS+S^2+S} \qquad (6)\end{aligned}$$

This classical estimator, the residual sum of squares divided by the number of data minus the number of independent parameters, is still biased contrary to many methods. It is unbiased only asymptotically. This statement has been confirmed by simulations.

2.2 Bayesian points of view

Regularised PCA has been presented and defined via the minimisation of the MSE in section 2.1. However, it is possible to define the method without any reference to MSE but using Bayesian considerations. It is well known in regression that there is a link between ridge regression and a Bayesian treatment of the regression model. More precisely, the maximum a posteriori of the regression parameters assuming a Gaussian prior for these parameters correspond to the ridge estimators (Hastie et al, 2009, p. 64). In this section we make the link between regularised PCA and Bayesian treatments of PCA.

2.2.1 Probabilistic PCA model

The probabilistic PCA (pPCA) model (Roweis, 1998; Tipping and Bishop, 1999) is a particular case of factor analysis model (Bartholomew, 1987) with an isotropic noise:

$$\begin{aligned}&\mathbf{x}_i = \mathbf{B}_{p\times S}\mathbf{z}_i + \varepsilon_i \\ &\qquad \mathbf{z}_i \sim \mathcal{N}(0,\mathbb{I}_S),\ \varepsilon_i \sim \mathcal{N}(0,\sigma^2\mathbb{I}_p)\end{aligned}$$

with $\mathbf{B}_{p\times S}$ the matrix of unknown coefficients and $\mathbb{I}_S$ and $\mathbb{I}_p$ the identity matrices of size S and p. This model induces a Gaussian distribution on the individuals with a specific structure of variance-covariance:

$$\mathbf{x}_i \sim \mathcal{N}(0,\mathbf{\Sigma}) \text{ with } \mathbf{\Sigma} = \mathbf{BB}' + \sigma^2\mathbb{I}_p$$

There is an explicit solution for the maximum likelihood estimators:

$$\hat{\mathbf{B}} = \mathbf{V}(\mathbf{\Lambda}-\sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{R} \qquad (7)$$

with $\mathbf{V}$ defined as in equation (2) *i.e.* as the matrix of the first S left singular vectors of $\mathbf{X}$, $\mathbf{\Lambda}$ a diagonal matrix constituted of the first S eigenvalues, $\mathbf{R}_{S\times S}$ a rotation matrix (usually equal to $\mathbb{I}_S$) and σ^2 estimated as the mean of the last eigenvalues.

This model can be seen as a random effect model by opposition to the fixed effect model (1), since the structure is random. It is more in agreement with cases where PCA is performed on sample data such as survey data. Indeed, the individuals are independent and identically distributed and only considered for the information they provide on the links between variables. In such studies, it does not make sense to consider "estimates" of the "individual parameters" since no parameter is associated with the individuals, only random variables ($\mathbf{z}_i$) are. However, estimators of the "individual parameters" are usually calculated as the expectation of the latent variables given the observed variables $\mathbb{E}(\mathbf{z}_i|\mathbf{x}_i)$. The calculation is detailed in Tipping and Bishop (1999) and leads to:

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}}+\sigma^2\mathbb{I}_S)^{-1} \qquad (8)$$

Such estimators are often called BLUP estimators (Robinson, 1991). Thus, using the maximum likelihood estimator of $\mathbf{B}$ (equation 7) and equation (8), it is possible to build a fit-

ted matrix as:

$$\begin{aligned}\hat{\mathbf{X}}^{\text{pPCA}} &= \hat{\mathbf{Z}}\hat{\mathbf{B}}' = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}} + \sigma^2\mathbb{I}_S)^{-1}\hat{\mathbf{B}}' \\ &= \mathbf{X}\mathbf{V}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{\Lambda}^{-1}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)^{\frac{1}{2}}\mathbf{V}' \\ &= \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S)\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}'\end{aligned}$$

since $\mathbf{X}\ \mathbf{V} = \mathbf{\Lambda}^{1/2}\mathbf{U}$ (given by the SVD of $\mathbf{X}$)

Therefore, considering the pPCA model leads to a fitted matrix of the same form than $\hat{\mathbf{X}}^{\text{rPCA}}$ presented in Section 2.1.2 with the same shrunk singular values $\left(\mathbf{\Lambda} - \sigma^2\mathbb{I}_S\right)\mathbf{\Lambda}^{-1/2}$. However, the pPCA model considers individuals as random, whereas they are fixed in model (1). Nevertheless, assuming a distribution on $\mathbf{z}_i$, which can be considered as the "individual parameters", is a way to define constraints on the individuals. Therefore, from a conceptual point of view, we prefer to consider this random effect model as a Bayesian treatment of the fixed effect model with a prior distribution on the left singular vectors.

Remark: Even if a maximum likelihood solution is available (equation 7), it is possible to use an EM algorithm (Rubin and Thayer, 1982) to estimate the parameters. The two steps are two multiple ridge regressions:

Step E: $\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}} + \hat{\sigma}^2\mathbb{I}_S)^{-1}$

Step M: $\hat{\mathbf{B}} = \mathbf{X}'\hat{\mathbf{Z}}(\hat{\mathbf{Z}}'\hat{\mathbf{Z}} + \hat{\sigma}^2\mathbf{\Lambda}^{-1})^{-1}$

The regularisation strategy can also be seen as the introduction of two ridge terms in the two following linear regressions leading to the usual PCA solution:

Step E: $\mathbf{U} = \mathbf{X}\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}$

Step M: $\mathbf{V} = \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}$

2.2.2 An empirical Bayesian approach

It is also possible to directly consider an empirical Bayesian treatment of the fixed effect model with a prior distribution on each cell of the data matrix per dimension: $\tilde{x}_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$. From model (1), this implies that $x_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2 + \frac{1}{\min(n-1;p)}\sigma^2)$.

The posterior distribution is:

$$p\left(\tilde{x}_{ij}^{(s)}|x_{ij}^{(s)}\right) = \frac{p\left(x_{ij}^{(s)}|\tilde{x}_{ij}^{(s)}\right)p\left(\tilde{x}_{ij}^{(s)}\right)}{p\left(x_{ij}^{(s)}\right)}$$

$$= \frac{\frac{1}{\sqrt{2\pi\frac{1}{\min(n-1;p)}\sigma^2}}\exp\left[-\frac{\left(x_{ij}^{(s)}-\tilde{x}_{ij}^{(s)}\right)^2}{2\frac{1}{\min(n-1;p)}\sigma^2}\right]\times\frac{1}{\sqrt{2\pi\tau_s^2}}\exp\left[-\frac{\left(\tilde{x}_{ij}^{(s)}\right)^2}{2\tau_s^2}\right]}{\frac{1}{\sqrt{2\pi(\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2)}}\exp\left[-\frac{\left(x_{ij}^{(s)}\right)^2}{2(\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2)}\right]}$$

$$= \frac{1}{\sqrt{2\pi\frac{\frac{1}{\min(n-1;p)}\sigma^2\tau_s^2}{\frac{1}{\min(n-1;p)}\sigma^2+\tau_s^2}}}\exp\left[-\frac{\left(\tilde{x}_{ij}^{(s)}-\frac{\tau_s^2}{\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2}x_{ij}^{(s)}\right)^2}{2\frac{\frac{1}{\min(n-1;p)}\sigma^2\tau_s^2}{\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2}}\right]$$

The expectation of the posterior distribution is obtained combining the likelihood and the priors:

$$\mathbb{E}\left(\tilde{x}_{ij}^{(s)}|x_{ij}^{(s)}\right) = \Phi_s x_{ij}^{(s)}$$

$$\text{with } \Phi_s = \frac{\tau_s^2}{\tau_s^2+\frac{1}{\min(n-1;p)}\sigma^2}$$

This expectation depends on unknown quantities. They are estimated by maximising the likelihood of $\left(x_{ij}^{(s)}\right)_{i=1,\ldots,n;j=1,\ldots,p}$ as a function of τ_s^2 to obtain:

$$\hat{\tau_s}^2 = \left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)$$

Consequently the shrinkage factor is estimated as $\hat{\Phi}_s = \frac{\left(\frac{1}{np}\lambda_s-\frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)}{\frac{1}{np}\lambda_s} = \frac{\lambda_s-\frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$ and also corresponds to the regularisation term defined in Section 2.1.1.

Remark: Hoff (2007) also proposed a Bayesian treatment of SVD-related models. His aim was, among others, to estimate the number of underlying dimensions. Roughly, his proposition consists in putting prior distributions on $\mathbf{U}$, $\mathbf{\Lambda}$, *and* $\mathbf{V}$. *More precisely, he uses von Mises uniform (Hoff, 2009) prior for orthonormal matrices (on the Steifeld manifold (Chikuse, 2003)) for* $\mathbf{U}$ *and* $\mathbf{V}$ *and normal priors for the singular values. It leads to a prior distribution for the structure* $\tilde{\mathbf{X}}$. *Then he builds a Gibbs sampler to get drawn from the posterior distributions. The posterior expectation of* $\tilde{\mathbf{X}}$ *can be used as a punctual estimate. It can also be seen as a regularised version of the maximum likelihood estimate. However, contrary to the previously exposed approach, there is no close form expression for the regularisation.*

2.3 Trade-off bias-variance

The rationale behind rPCA can be illustrated on graphical representations. Usually, different types of graphical representations are associated with PCA depending on whether the left and right singular vectors are represented as normed to 1 or to their associated singular value. In our practice, we represent the individual coordinates by $\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$ and the variable coordinates by $\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$. Therefore, the global shape of the individual cloud represents the variance. Similarly, in the variable representation, the cosine of the angle between two variables can be interpreted as the covariance. Since rPCA comes down to modify the singular values, it will affect both the representation of the individuals and of the variables. We focus here on the individual representation.

Data are generated according to model (1) with an underlying signal $\tilde{\mathbf{X}}_{5\times15}$ composed of 5 individuals and 15 variables in two dimensions. Then, 300 matrices are generated with the same underlying structure: $\mathbf{X}^{sim} = \tilde{\mathbf{X}}_{5\times15}+$

ε_{sim} with $sim = 1, ..., 300$. On each data matrix, PCA and rPCA are performed. In figure 1, the configurations of the 5 individuals obtained after each PCA are represented on the left, whereas the configurations obtained after each rPCA, are represented on the right. The average configurations over the 300 simulations are represented by triangles and the true individual configuration obtained from the PCA of $\tilde{\mathbf{X}}$ is represented by big dots. Representing several sets of coordinates from different PCAs can suffer from translation, reflection, dilatation or rotation ambiguities. That is why all configurations are superimposed using Procustes rotations (Gower and Dijksterhuis, 2004) by taking as the reference the true individual configuration.

Compared to PCA, rPCA provides a representation which is more biased because the coordinates of the average points (triangles) are systematically inferior to the coordinates of the true points (big dots). This is expected because the regularisation term shrinks the individual coordinates towards the origin. In addition, as it is clear for individual number 4 (dark blue), the representation is less variable. Figure 1 gives thus a rough idea of the trade-off bias-variance. Note that even the PCA representation is biased, but this is also expected since $\mathbb{E}(\hat{\mathbf{X}}) = \tilde{\mathbf{X}}$ only asymptotically as detailed section 2.1.1.

3 Simulation study

To assess rPCA, a simulation study is conducted and rPCA is confronted to classical PCA as well as the SURE method proposed by Candès et al (2012). As explained in the introduction, the SURE method relies on a soft thresholding strategy:

$$\hat{x}_{ij}^{\text{SURE}} = \sum_{s=1}^{\min(n,p)} \left(\sqrt{\lambda_s} - \lambda\right)_+ u_{is} v_{js},$$

The shrinkage parameter λ is automatically selected by minimising Stein's Unbiased Risk Estimate (SURE). As a tuning parameter, the SURE method does not require the number of underlying dimensions of the signal, but it does require to estimate the noise variance σ^2 to determine λ.

3.1 Recovery of the signal

Data are simulated according to model (1). The structure is simulated by varying several parameters:

- the number of individuals n and the number of variables p based on 3 different pairs: ($n = 100$ and $p = 20$; $n = 50$ and $p = 50$; $n = 20$ and $p = 100$)
- the number of underlying dimensions S (2; 4)
- the ratio of the first eigenvalue on the second eigenvalue (d_1/d_2) of $\tilde{\mathbf{X}}$ (4; 1). When the number of underlying dimensions is higher than 2, the subsequent eigenvalues are roughly of the same order of magnitude.

More precisely, $\tilde{\mathbf{X}}$ is simulated as follows:

1. A SVD is performed on a $n \times S$ matrix generated from a standard multivariate normal distribution. The left singular vectors provide S empirically orthonormal vectors.
2. Each vector $s : 1, ..., S$ is replicated to obtain the p variables. The number of times that each vector s is replicated depends on the ratio between the eigenvalues (d_1/d_2). For instance, if $p = 50$, $S = 2$, $(d_1/d_2) = 4$, the first vector is replicated 40 times and the second vector is replicated 10 times.

Then, to generate the matrix $\mathbf{X}$, a Gaussian isotropic noise is added to the structure. Different levels of variance σ^2 are considered to obtain three signal-to-noise ratios (Mazumder et al, 2010) equal to 4, 1 and 0.8. A high signal-to-noise ratio (SNR) implies that the variables of $\mathbf{X}$ are very correlated, whereas a low SNR implies that the data are very noisy. For each combination of the parameters, 500 data sets are generated.

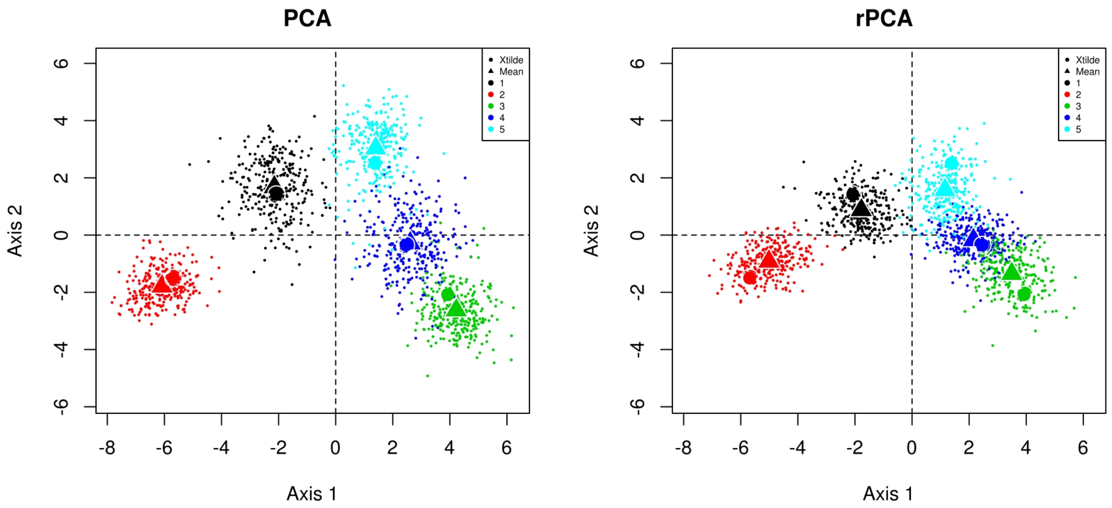


Fig. 1: Superimposition of several configurations of individual coordinates using Procustes rotations towards the true individual configuration of $\tilde{\mathbf{X}}_{5\times 15}$ (big dots). Configurations of the PCA (left) and the rPCA (right) of each $\mathbf{X} = \tilde{\mathbf{X}} + \varepsilon_{sim}$, with $sim = 1, ..., 300$ are represented with small dots. The average configuration over the 300 configurations is represented by triangles.

To assess the recovery of the signal, the MSE is calculated between $\hat{\mathbf{X}}$ obtained from each method and $\tilde{\mathbf{X}}$. The fitted matrices obtained by PCA and rPCA are obtained considering the true number of underlying dimensions as known. The SURE method is performed with the true noise variance as in Candès et al (2012). Results of the simulation study are gathered in Table 1.

First, rPCA outperforms both PCA and the SURE method in almost all the situations. As expected, the MSE obtained by PCA and rPCA are roughly of the same order of magnitude when the SNR is high (SNR = 4) whereas rPCA is better than PCA when data are noisy (SNR = 0.8). The differences between rPCA and PCA are also more important when the ratio (d_1/d_2) is high than when the eigenvalues are equal. When (d_1/d_2) is large, the signal is concentrated on the first dimension whereas it is scattered in more dimensions when the ratio is smaller. Consequently, a same amount of noise has a greater impact on the second dimension in the first case. This may increase the advantage of rPCA which tends to reduce the impact of noise.

The main characteristic of the SURE method observed in all the simulations is that it gives particularly good results when the data are very noisy. Consequently, the results are satisfactory when SNR = 0.8 and this is all the more true than the number of underlying dimensions is high. This behaviour can be explained by the fact that a same amount of signal is more impacted by the by the noise if the signal is scattered on many dimensions. This remark highlights the fact that the SNR is not necessarily a good measure of the level of noise in a data set. In addition, the results of the SURE method are quite poor when the SNR is high. This can be explained by the fact that the SURE method takes into account too many dimensions (since all the singular values greater than the threshold are kept) in the estimation of $\hat{\mathbf{X}}^{\text{SURE}}$. For example, with $n = 100$, $p = 20$, $S = 2$, $SNR = 4$ and $(d_1/d_2) = 4$ (first row of Table 1), the SURE method considers between 9 and 13 dimensions to estimate $\hat{\mathbf{X}}^{\text{SURE}}$.

Finally, the behaviour regarding the ratio (n/p) is worth commenting. The MSEs are in

Table 1: Mean Squared Error (and its standard deviation) between $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ for PCA, rPCA and SURE method over 500 simulations. Results are given for different numbers of individuals (n), numbers of variables (p), numbers of underlying dimensions (S), signal-to-noise ratios (SNR) and ratios of the first eigenvalue on the second eigenvalue (d_1/d_2).

n	p	S	SNR	(d_1/d_2)	$MSE(\hat{\mathbf{X}}^{\text{PCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{rPCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\text{SURE}}, \tilde{\mathbf{X}})$
100	20	2	4	4	4.22E-04 (1.69E-06)	**4.22E-04** (1.69E-06)	8.17E-04 (2.67E-06)
100	20	2	4	1	4.21E-04 (1.75E-06)	**4.21E-04** (1.75E-06)	8.26E-04 (2.89E-06)
100	20	2	1	4	1.26E-01 (5.29E-04)	**1.08E-01** (4.56E-04)	1.60E-01 (6.15E-04)
100	20	2	1	1	1.23E-01 (5.05E-04)	**1.11E-01** (4.61E-04)	1.69E-01 (6.28E-04)
100	20	2	0.8	4	3.34E-01 (1.38E-03)	**2.40E-01** (9.90E-04)	3.10E-01 (1.05E-03)
100	20	2	0.8	1	3.12E-01 (1.38E-03)	**2.45E-01** (1.10E-03)	3.32E-01 (1.22E-03)
100	20	4	4	4	8.25E-04 (2.39E-06)	**8.24E-04** (2.38E-06)	1.42E-03 (3.54E-06)
100	20	4	4	1	8.26E-04 (2.38E-06)	**8.25E-04** (2.38E-06)	1.43E-03 (3.48E-06)
100	20	4	1	4	2.60E-01 (8.44E-04)	**1.96E-01** (6.51E-04)	2.43E-01 (6.84E-04)
100	20	4	1	1	2.47E-01 (7.16E-04)	**2.04E-01** (5.99E-04)	2.62E-01 (6.94E-04)
100	20	4	0.8	4	7.41E-01 (2.69E-03)	**4.27E-01** (1.53E-03)	4.36E-01 (1.11E-03)
100	20	4	0.8	1	6.68E-01 (2.02E-03)	**4.40E-01** (1.40E-03)	4.83E-01 (1.33E-03)
50	50	2	4	4	2.81E-04 (1.32E-06)	**2.81E-04** (1.32E-06)	5.95E-04 (2.24E-06)
50	50	2	4	1	2.79E-04 (1.24E-06)	**2.79E-04** (1.24E-06)	5.93E-04 (2.21E-06)
50	50	2	1	4	8.48E-02 (4.09E-04)	**7.82E-02** (3.85E-04)	1.26E-01 (4.97E-04)
50	50	2	1	1	8.21E-02 (3.87E-04)	**7.77E-02** (3.70E-04)	1.31E-01 (5.08E-04)
50	50	2	0.8	4	2.30E-01 (1.12E-03)	**1.93E-01** (9.64E-04)	2.55E-01 (1.01E-03)
50	50	2	0.8	1	2.14E-01 (9.58E-04)	**1.89E-01** (8.57E-04)	2.73E-01 (1.07E-03)
50	50	4	4	4	5.48E-04 (1.84E-06)	**5.48E-04** (1.84E-06)	1.04E-03 (2.82E-06)
50	50	4	4	1	5.46E-04 (1.76E-06)	**5.46E-04** (1.76E-06)	1.04E-03 (2.79E-06)
50	50	4	1	4	1.75E-01 (6.21E-04)	**1.53E-01** (5.54E-04)	2.00E-01 (5.79E-04)
50	50	4	1	1	1.68E-01 (5.49E-04)	**1.52E-01** (5.08E-04)	2.09E-01 (6.04E-04)
50	50	4	0.8	4	5.07E-01 (1.90E-03)	3.87E-01 (1.53E-03)	**3.85E-01** (1.12E-03)
50	50	4	0.8	1	4.67E-01 (1.62E-03)	**3.76E-01** (1.38E-03)	4.13E-01 (1.23E-03)
20	100	2	4	4	4.22E-04 (1.72E-06)	**4.22E-04** (1.72E-06)	8.15E-04 (2.80E-06)
20	100	2	4	1	4.21E-04 (1.69E-06)	**4.20E-04** (1.70E-06)	8.20E-04 (2.89E-06)
20	100	2	1	4	1.25E-01 (5.35E-04)	**1.06E-01** (4.53E-04)	1.57E-01 (5.83E-04)
20	100	2	1	1	1.22E-01 (5.28E-04)	**1.10E-01** (4.76E-04)	1.67E-01 (6.20E-04)
20	100	2	0.8	4	3.30E-01 (1.43E-03)	**2.35E-01** (1.03E-03)	3.06E-01 (1.13E-03)
20	100	2	0.8	1	3.18E-01 (1.30E-03)	**2.50E-01** (1.03E-03)	3.34E-01 (1.25E-03)
20	100	4	4	4	8.28E-04 (2.38E-06)	**8.27E-04** (2.39E-06)	1.41E-03 (3.64E-06)
20	100	4	4	1	8.29E-04 (2.58E-06)	**8.28E-04** (2.58E-06)	1.42E-03 (3.68E-06)
20	100	4	1	4	2.55E-01 (7.59E-04)	**1.97E-01** (5.92E-04)	2.45E-01 (6.47E-04)
20	100	4	1	1	2.48E-01 (7.45E-04)	**2.04E-01** (6.20E-04)	2.60E-01 (6.91E-04)
20	100	4	0.8	4	7.13E-01 (2.55E-03)	**4.15E-01** (1.47E-03)	4.37E-01 (1.19E-03)
20	100	4	0.8	1	6.66E-01 (2.01E-03)	**4.34E-01** (1.31E-03)	4.78E-01 (1.24E-03)

the same order of magnitude for $(n/p) = 0.2$ and $(n/p) = 5$ and much smaller for $(n/p) = 1$ for all the methods. The issue of dimensionality does not lie in the fact that the number of variables is much larger than the number individuals. However, difficulties are encountered when either one mode (n or p) is larger than the other one which can be explained by the bilinear form of the model.

3.2 Simulations from Candès et al (2012)

Regularised PCA is also assessed on the simulations of Candès et al (2012). Simulated matrices of size 200×500 were drawn with 4 SNR (0.5, 1, 2 and 4) and 2 numbers of underlying dimensions (10, 100).

Results for the SURE method (Table 2) are in agreement with the results obtained by Candès et al (2012). As in the first simulation study (section 3.1), rPCA outperforms both PCA and the SURE method in almost all cases. However, the SURE method provides better results than rPCA when the number of underlying dimensions S is high ($S = 100$) and the SNR is small (SNR = 1, 0.5). This is in agreement with the previous comments highlighting the ability of the SURE method to handle noisy situations. Nevertheless, we can note that when the SNR is equal to 0.5, rPCA

is performed with the "true" number of underlying dimensions (100) whereas estimating the number of underlying dimensions would lead, on this data, to conserve 0 dimensions, according to several methods. Data are very noisy and the signal can be considered as nearly lost. Results obtained with rPCA, taking into account 0 dimension, boils down to predict all the values by 0 which corresponds to an MSE equal to 1. In this case, considering 0 dimensions in rPCA leads to a lower MSE than taking into account 100 dimensions (MSE = 1.48), but it is still higher than the MSE of the SURE method (0.85).

The R (R Core Team, 2012) code to perform all the simulations is available on the web page of the authors.

3.3 Recovery of the graphical outputs

Our objective is also to obtain graphical outputs (individual and variable representations) which are as close as possible to the outputs obtained from $\tilde{\mathbf{X}}$. Let us consider a toy data set with 100 individuals, 20 variables, 2 underlying dimensions, $(d_1/d_2) = 4$ and a SNR equal to 0.8 (row 5 of Table 1).

Figure 2 provides the true individual representation obtained from $\tilde{\mathbf{X}}$ (top left) as well as the representations obtained by PCA (top right), rPCA (bottom left) and the SURE method (bottom right). The cloud associated with PCA has a higher variability than the cloud associated with rPCA which is tightened around the origin. The effect of regularisation is stronger on the second axis than on the first one, which is expected because of the regularisation term. For instance, the individuals 82 and 59 which have small coordinates on the second axis in PCA are brought closer to the origin in the representation obtained by rPCA which is more in agreement with the true configuration. The cloud associated with the SURE method is tightened around the origin on the first axis and even more on the second one, which is also expected because of the regularisation term. However the global variance of the SURE representation, which is reflected by the variability, is clearly lower than the variance of the true signal. Therefore, the global shape of the cloud of rPCA is the closest to the true one and thus rPCA recovers well the distances between individuals.

Figure 3 provides the corresponding representations for the variables. The link between the variables which have high coordinates on the first and the second axis of the PCA of $\mathbf{X}$ is reinforced in rPCA. This is consistent with the representation of $\tilde{\mathbf{X}}$. For instance, variables 9 and 7 which are correlated to 1 in $\tilde{\mathbf{X}}$ are not very linked in the PCA representation (correlation equal to 0.68) whereas their correlation equals 0.81 in the rPCA representation and 0.82 in the SURE representation. On the contrary, variables 20 and 7, orthogonal in $\tilde{\mathbf{X}}$, have rather high coordinates, in absolute value, on the second axis in the PCA representation (correlation equal to -0.60). Their link is slightly weakened in the rPCA representation (correlation equal to -0.53) and in the SURE representation (correlation equal to -0.51). In addition, all the variables are generated with a variance equal to 1. The variances are over-estimated in the PCA representation and under-estimated in the SURE representation, particularly for the variables which are highly linked to the second axis. The best compromise for the variances is provided by rPCA. Therefore, rPCA recovers well the variances and the covariances of the variables.

4 Application to real data sets: transcriptome profiling

Regularised PCA is applied to a real data set (Désert et al, 2008) which consists in a collection of 12664 gene expressions in 27 chickens submitted to 4 nutritional statuses: continuously fed (N), fasting for 16 hours (F16), fasting for 16 hours then refed for 5 hours (F16R5), fasting for 16 hours then refed for 16 hours (F16R16).

Table 2: Mean Squared Error (and its standard deviation) between $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$ for PCA, regularised PCA (rPCA) and SURE method over 100 simulations. Results are given for $n = 200$ individuals, $p = 500$ variables, different numbers of underlying dimensions (S) and signal-to-noise ratios (SNR).

S	SNR	$MSE(\hat{\mathbf{X}}^{\mathrm{PCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\mathrm{rPCA}}, \tilde{\mathbf{X}})$	$MSE(\hat{\mathbf{X}}^{\mathrm{SURE}}, \tilde{\mathbf{X}})$
10	4	4.31E-03 (7.96E-07)	**4.29E-03** (7.91E-07)	8.74E-03 (1.15E-06)
10	2	1.74E-02 (2.84E-06)	**1.71E-02** (2.81E-06)	3.29E-02 (4.68E-06)
10	1	7.16E-02 (1.25E-05)	**6.75E-02** (1.15E-05)	1.16E-01 (1.59E-05)
10	0.5	3.19E-01 (5.44E-05)	**2.57E-01** (4.85E-05)	3.53E-01 (5.42E-05)
100	4	3.79E-02 (2.02E-06)	**3.69E-02** (1.93E-06)	4.50E-02 (2.12E-06)
100	2	1.58E-01 (8.99E-06)	**1.41E-01** (7.98E-06)	1.56E-01 (8.15E-06)
100	1	7.29E-01 (4.84E-05)	4.91E-01 (2.96E-05)	**4.48E-01** (2.26E-05)
100	0.5	3.16E+00 (1.65E-04)	1.48E+00 (1.12E-04)	**8.52E-01** (3.07E-05)

Since there are 4 nutritional statuses, 3 dimensions are considered. We expect the first three principal components to represent the between class variability, whereas the following components represent the within class variability which is less of interest. Figure 4 shows the individual representations obtained by PCA (top left), rPCA (top right) and the SURE method (bottom left). To better highlight the effect of regularisation, the dimensions 1 and 3 are represented. The first dimension of PCA, rPCA and the SURE method order the nutritional statuses from the continuously fed chickens (left) to the fasting chickens (right). The chicken N.4 and the chicken F16R5.1 which have high coordinates, in absolute value, on the third axis of PCA are brought closer to the other chickens submitted to the same status in the rPCA representation and in the SURE representation. In addition, the chickens N.1 and F16.4 which have high coordinates on the first axis are brought closer to the origin in the SURE representation. Despite these differences, the impact of the regularisation on the graphical outputs appears to be small.

The representation obtained after a sparse PCA (sPCA) method (Witten et al, 2009) implemented in the R package `PMA` (Witten et al, 2011) is also provided (bottom right). Indeed, it is very common to use sparse methods on this kind of data (Zou et al, 2006). The basic assumptions for the development of sPCA is that PCA provides principal components that are linear combinations of the original variables which may lead to difficulties during the interpretation especially when the number of variables is huge. Loadings obtained via sPCA are indeed sparse which means they contain many 0 elements which comes down to select variables. The representation stemming from sPCA is quite different from the other representations, particularly the clusters of F16R5 and of F16 chickens are less clearly differentiated.

It is usual to complement principal components methods with heatmaps (Eisen et al, 1998) in order to simultaneously cluster the chickens and the genes. Because rPCA modifies the distances between chickens as well as the covariances between genes, the rPCA heatmap will differ from the PCA heatmap. The heatmap clustering is applied to the matrices $\hat{\mathbf{X}}$ obtained by the different methods (Figure 5). The rPCA heatmap (Figure 5b) is much more satisfying than the PCA heatmap (Figure 5a). Indeed, the chickens submitted to 16 hours of fasting are distinguished into two subclusters in the PCA heatmap separated by the chickens F16R5.1, F16R16.3 and F16R16.4, whereas they are well-clustered in the rPCA heatmap. Similarly the chickens F16R5 are gathered in the PCA heatmap except for chickens F16R5.1 and F16R5.3, whereas they are well-clustered in the rPCA heatmap. Finally,

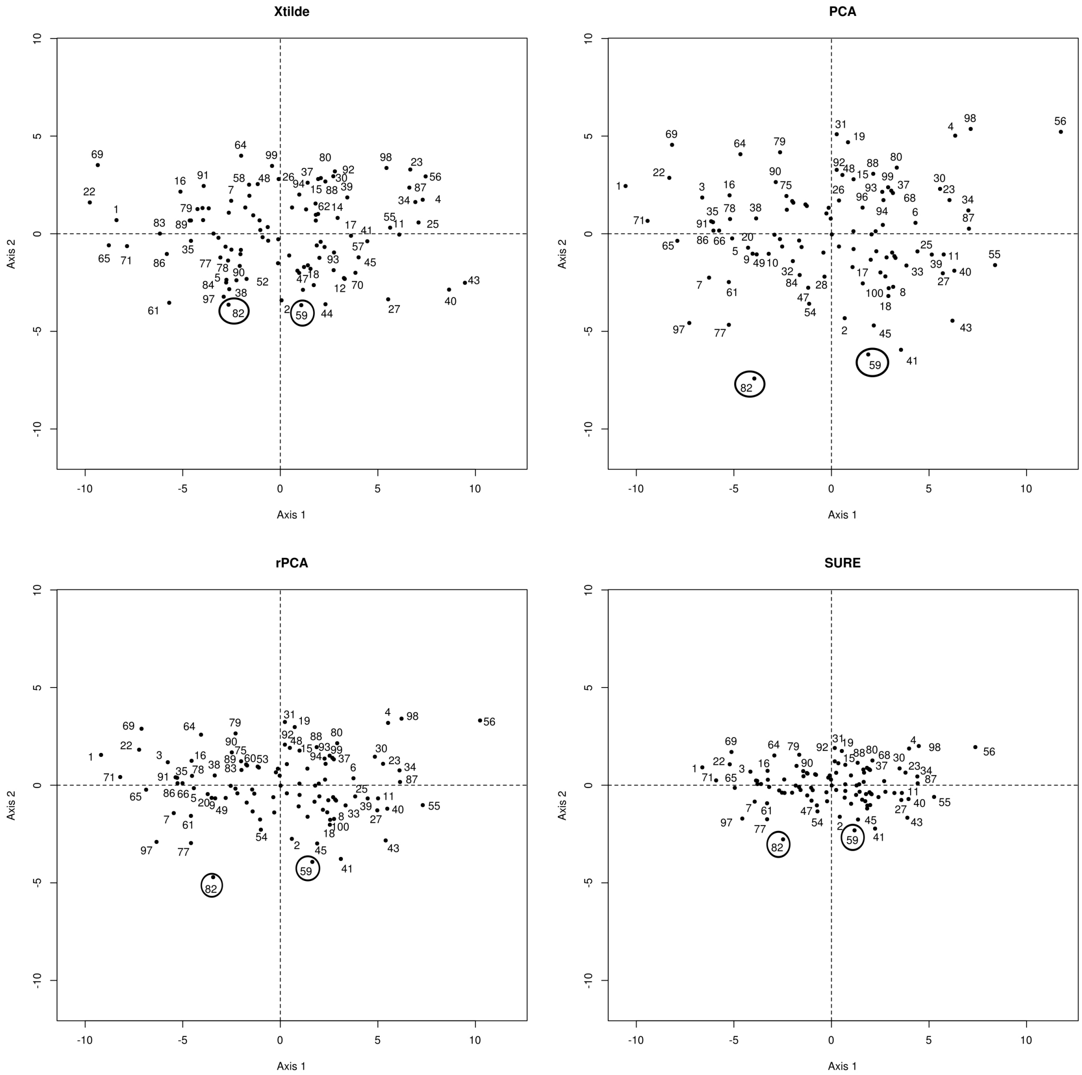


Fig. 2: Individual representations of $\tilde{\mathbf{X}}$ (top left), of the PCA of $\mathbf{X}$ (top right), of the rPCA of $\mathbf{X}$ (bottom left) and of the SURE method applied to $\mathbf{X}$ (bottom right) for a data set with $n = 100$, $p = 20$, $S = 2$, $(d_1/d_2) = 4$ and SNR$= 0.8$.

the F16R16 chickens are more scattered in both representations. However in rPCA, it can be interpreted as some of the chickens having fully recovered from the fasting period are mixed with continuously fed chickens and some having not fully recovered are mixed with F16R5 chickens: the large majority of F16R16 chickens are gathered and mixed with N.6 and N.7, and chicken F16R16.1 is mixed with F16R5 chickens. It is not the case for PCA, where the F16R16 chickens are mixed with chickens submitted to all the other nutritional statuses. The conclusions concerning the SURE heatmap (Figure 5c) are roughly similar to the conclusions drawn for rPCA. Globally, the 4 clusters corresponding to the 4 nutritional statuses are well-defined. However, the chicken F16R5.3 is clustered with the N chickens. In addition, the

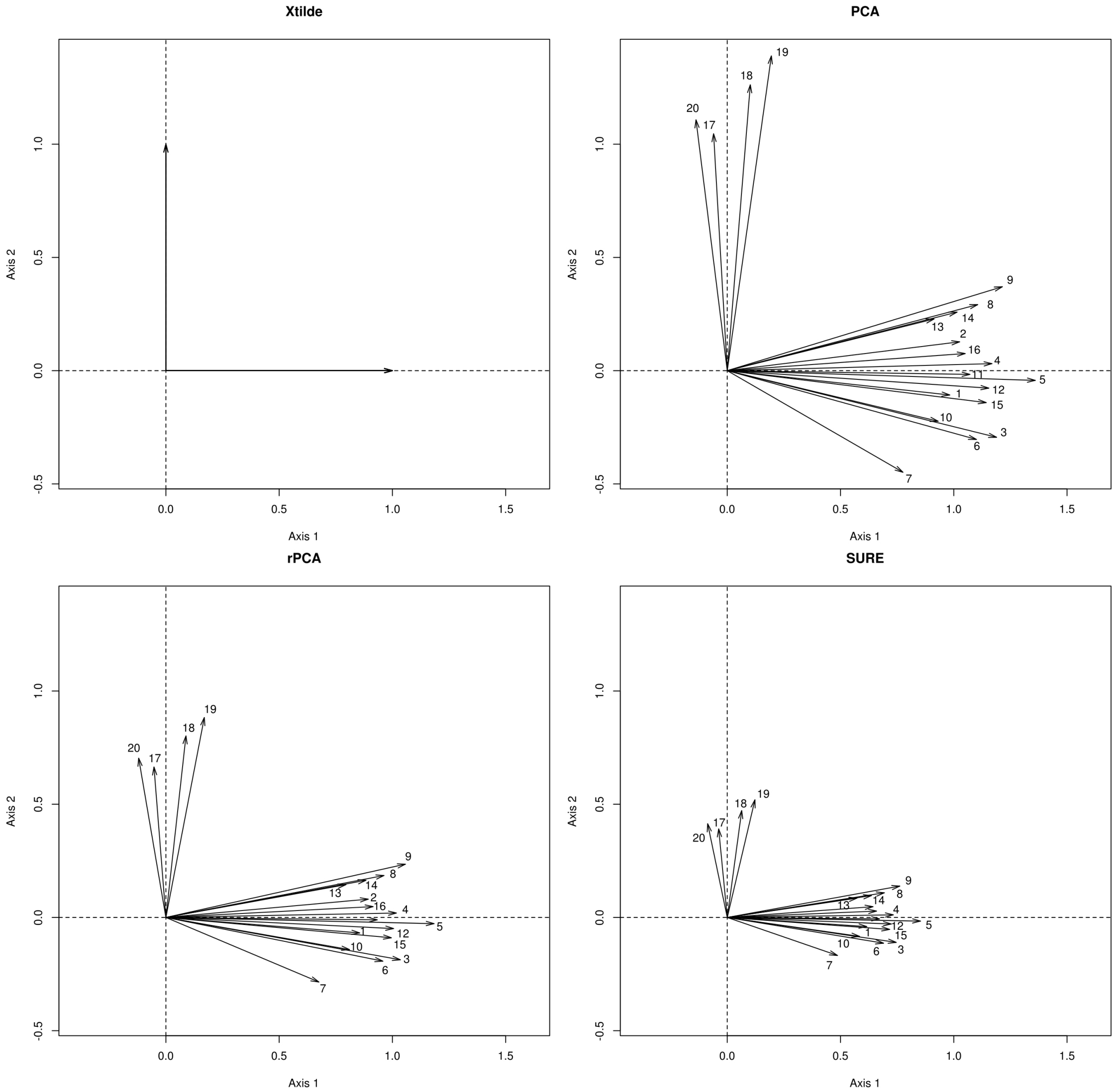


Fig. 3: Variable representations of the PCA of $\tilde{\mathbf{X}}$ (top left), the PCA of $\mathbf{X}$ (top right), the rPCA of $\mathbf{X}$ (bottom left) and the SURE method applied to $\mathbf{X}$ (bottom right) for an example of data set with $n = 100$, $p = 20$, $S = 2$, $(d_1/d_2) = 4$ and SNR$= 0.8$.

global contrasts are weaker in the SURE heatmap than in the rPCA heatmap. The heatmap stemming from sPCA (Figure 5d) seems to be easier to interpret (more contrasts), due to the drastic selection of genes (43 on the 12664). However none of the chicken clusters is clearly defined.

We will not dwell on the interpretation of the gene expressions in the heatmap, however, if the chicken clustering is coherent, the gene clustering is expected to be more coherent as well.

Conclusion

When data can be seen as a true signal marred by errors, PCA does not provide the best re-

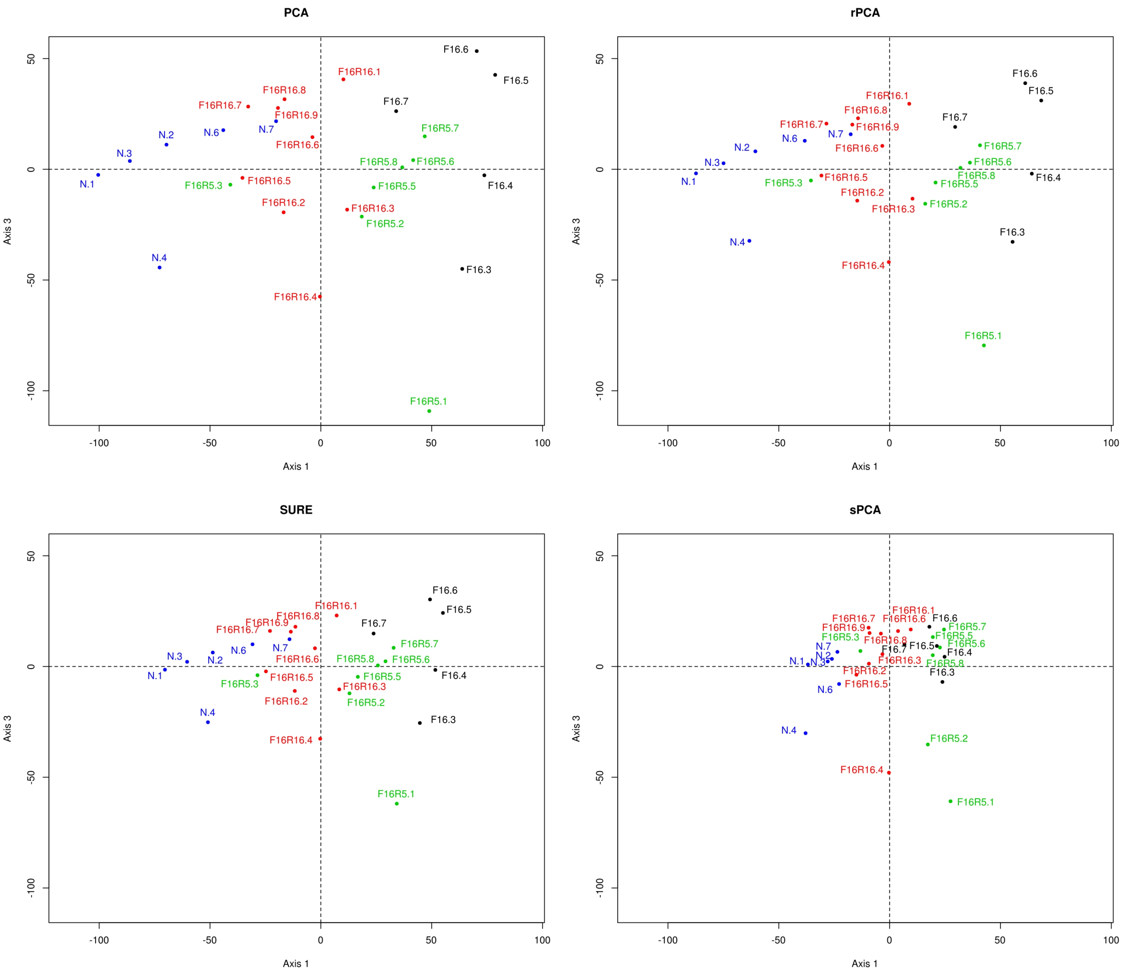
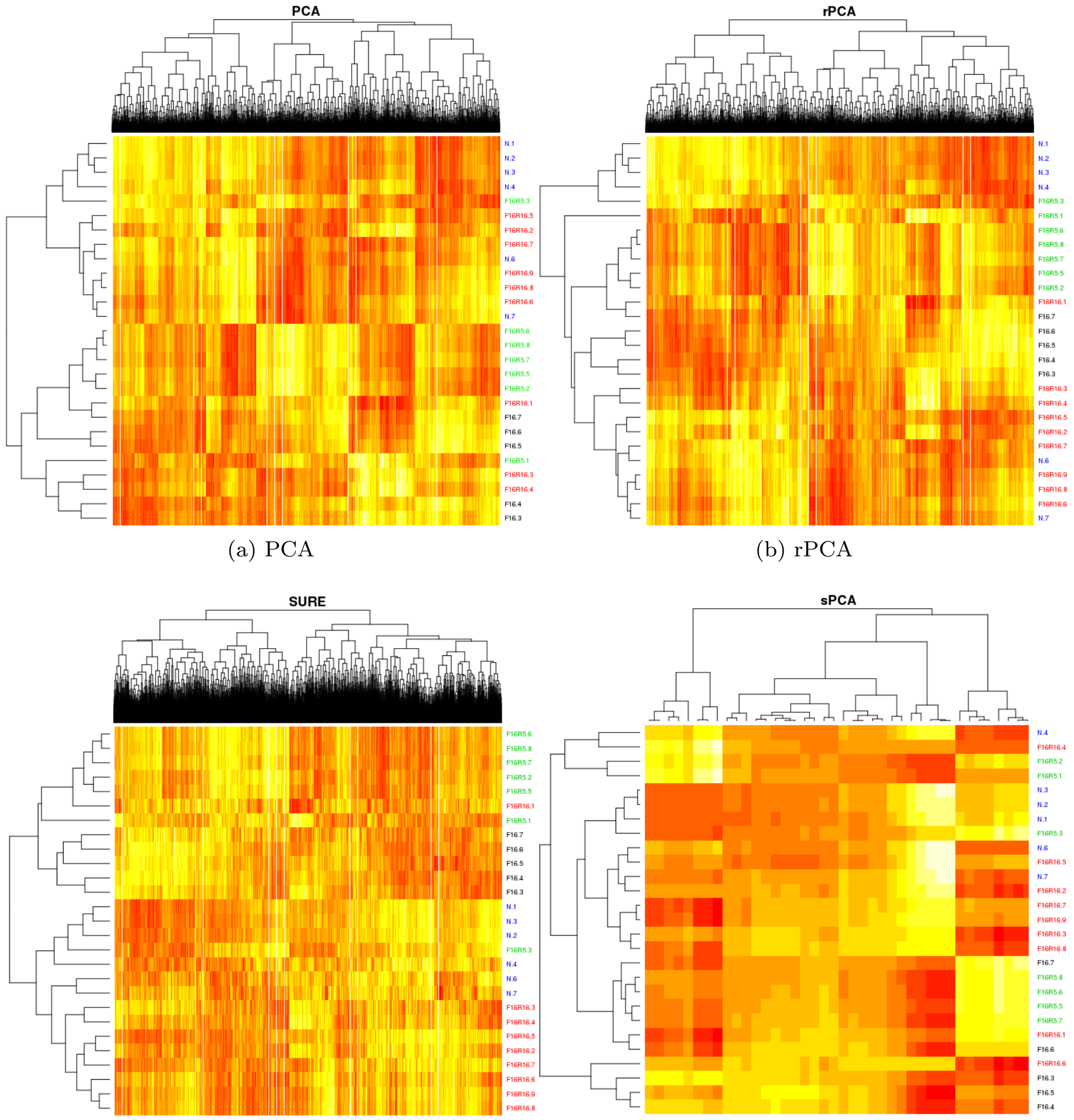


Fig. 4: Representation of the individuals on the dimension 1 and 3 of the PCA (top left), the rPCA (top right), the SURE method (bottom left) and sPCA (bottom right) of the transcriptome profiling data. Individuals are coloured according to the nutritional statuses.

covery of the underlying signal. Thresholding the singular values improves the estimation of the underlying structure especially when data are noisy. Soft thresholding is one of the most popular strategies and consists in linearly shrinking the singular values. In this framework,
Candès et al (2012) developed an automatic strategy to determine the regularisation parameter. The regularised PCA proposed in this paper applies a nonlinear transformation of the singular values and a hard thresholding rule. The regularised term is analytically derived from the MSE using asymptotic results from nonlinear regression models. In the simulations, rPCA outperforms the SURE method in most of the situations.

However, rPCA requires a tuning parameter which is the number of underlying dimensions. Many methods are available in the literature to select this parameter. However, it is still a difficult problem and an active research area. A classical statement is the following one: if the selected number of dimensions is smaller



(a) PCA (b) rPCA

(c) SURE (d) sPCA

Fig. 5: Heatmaps associated with the analysis of the transcriptomic data set. The data sets used to perform the heatmaps are the fitted matrices stemming from PCA (a), rPCA (b), the SURE method (c) and sPCA (d).

than the rank S of the signal, a part of the relevant information is lost and, in our situation, it results in overestimating the noise variance. On the contrary, selecting more than S dimensions seems better because all the signal is taken into account even if the noise variance is underestimated. However, the noise may overwhelm the signal which lead to “loose” the dimensions of the signal which have the smallest variability. In such a case, it is better to

select a number of dimensions smaller than S. This strategy is a way to regularise more which is acceptable when data are very noisy. In practice, we use a cross-validation strategy Josse and Husson (2011) which has a good behaviour in our simulations (find the true number of dimensions when the signal-to-noise ratio is large, and find a smaller number when the signal-to-noise ratio is small).

The graphical outputs provided by rPCA are closer to the ones that would be obtained from the signal only since rPCA improves the recovery of the signal compared to classical PCA. However, the graphical representations are biased. Another solution could be to keep the graphical representations obtained from PCA and to enrich them through confidence areas. Such areas may be constructed using a residual bootstrap procedure.

References

Bartholomew D (1987) Latent Variable Models and Factor Analysis. Charles Griffin and Co Ltd

Candès EJ, Tao T (2009) The power of convex relaxation: Near-optimal matrix completion. IEEE Trans Inform Theory 56(5):2053–2080

Candès EJ, Sing-Long CA, Trzasko JD (2012) Unbiased risk estimates for singular value thresholding and spectral estimators, (Submitted)

Caussinus H (1986) Models and uses of principal component analysis (with discussion), DSWO Press, p 149–178

Chikuse Y (2003) Statistics on Special Manifolds. Springer

Cornelius P, Crossa J (1999) Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. Crop Science 39:998–1009

Denis JB, Gower JC (1994) Asymptotic covariances for the parameters of biadditive models. Utilitas Mathematica pp 193–205

Denis JB, Gower JC (1996) Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. Journal of the Royal Statistical Society Series C (Applied Statistics) 45:479–493

Denis JB, Pázman A (1999) Bias of least squares estimators in nonlinear regression models with constraints. part ii: Biadditive models. Applications of Mathematics 44:359–374

Désert C, Duclos M, Blavy P, Lecerf F, Moreews F, Klopp C, Aubry M, Herault F, Le Roy P, Berri C, Douaire M, Diot C, S L (2008) Transcriptome profiling of the feeding-to-fasting transition in chicken liver. BMC Genomics

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95(25):14,863–14,868

Gower JC, Dijksterhuis GB (2004) Procrustes Problems. Oxford University Press, USA

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. Springer

Hoff P (2007) Model averaging and dimension selection for the singular value decomposition. J Amer Statist Assoc 102(478):674–685

Hoff PD (2009) Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. Journal of Computational and Graphical Statistics 18(2):438–456

Hwang H, Tomiuk M, Takane Y (2009) Correspondence Analysis, Multiple Correspondence Analysis and Recent Developments, Sage Publications, pp 243–263

Josse J, Husson F (2011) Selecting the number of components in pca using cross-validation approximations. Computational Statististics and Data Analysis 56:1869–1879

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research 99:2287–2322

Papadopoulo T, Lourakis MIA (2000) Estimating the jacobian of the singular value

decomposition: Theory and applications. In: In Proceedings of the European Conference on Computer Vision, ECCV00, Springer, pp 554–570

R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statistical Science 6(1):15–32

Roweis S (1998) Em algorithms for pca and spca. In: Advances in Neural Information Processing Systems, MIT Press, pp 626–632

Rubin DB, Thayer DT (1982) EM algorithms for ML factor analysis. Psychometrika 47(1):69–76

Takane Y, Hwang H (2006) Regularized multiple correspondence analysis, Boca Raton, FL: Chapman & Hall/CRC, pp 259–279

Tipping M, Bishop C (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society (Series B) 61:611–622

Witten D, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10:515–534

Witten D, Tibshirani R, Gross S, Narasimhan B (2011) PMA: Penalized Multivariate Analysis. URL `http://CRAN.R-project.org/package=PMA`, r package version 1.0.8

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2):265–286